Supplementary Materials for Rethinking the Region Classification in Open-Vocabulary Semantic Segmentation: An Image-to-Image View

Yuan Wang[1*] Rui Sun[1*] Naisong Luo[1] Yuwen Pan[1] Tianzhu Zhang[1,2†]
[1]University of Science and Technology of China
[2]Deep Space Exploration Laboratory
{wy2016, issunrui, lns6, panyw}@mail.ustc.edu.cn, {tzzhang}@ustc.edu.cn

In the supplementary material, we first introduce more implementation details of visual foundation models exploited in our experiments in Sec 1. Then in Sec 2 we conduct expanded comparative analyses between the proposed RIM and a line of methodologies on different datasets. Finally, we show more qualitative results of our method in Sec 3.

1. More Implementation Details

Details about image generation. To construct the intra-modal reference features, we resort to the Stable Diffusion [8] model to generate $K = 50$ images for each candidate category. We employ 50 denoising steps for reference image generation.

Details about prompt sampling. To avoid the concentration of multiple prompt points in a smaller area, which results in SAM only segmenting a part of the foreground, we follow [5] to employ a distance transform algorithm to uniformly sample prompt points on the binarized cross-attention map. In particular, we identify the points that are maximally distant from both the edge of the binarized cross-attention map and each other as the prompt points as described in Algorithm 1. Then we input these prompt points into SAM to facilitate the segmentation of the complete foreground in the synthesized image.

Details about mask proposals generation. SAM tends to generate multiple corresponding mask proposals from a single prompt point, and these mask proposals may exhibit hierarchical relationships, such as "*T-shirt*" and "*person*". To acquire more complete masks for instances, we perform a mask fusing on the mask proposals generated by SAM. Specifically, we merge two mask proposals M_1 and M_2 if

$$\frac{\mathrm{Area}(\mathrm{M}_1 \cap \mathrm{M}_2)}{\min(\mathrm{Area}(\mathrm{M}_1), \mathrm{Area}(\mathrm{M}_2)} > 0.9, \tag{1}$$

where the Area means the measure of the area of a mask proposal. Mask fusing can prevent overly fragmented mask proposals from impacting the performance of region classification.

Algorithm 1 The Process of Prompt Points Sampling

Input: The binarized cross-attention map M, number of prompt points N to sample. Set X represents the locations of foreground pixels, and the set L containing both background pixel locations (B) and sampled prompt points (P);

Output: Set of positions of prompt points $\boldsymbol{P} = \{p_i\}_{n=1}^{N}$;

1: Initializing the set $\boldsymbol{P} = \{\}$;
2: **for** n in $\{1, 2, \ldots, N\}$ **do**
3: Calculating the distance transform between locations $\boldsymbol{y} \in \boldsymbol{L}$ and a specific foreground location $\boldsymbol{x} \in \boldsymbol{X}$:

$$D(\boldsymbol{x}) = \min_{\boldsymbol{y} \in \boldsymbol{L}} \|\boldsymbol{x} - \boldsymbol{y}\|_2 ;$$

4: Selecting the furthest distance $\boldsymbol{p}^*$:

$$\boldsymbol{p}^* = \arg\max_{\boldsymbol{x}} D(\boldsymbol{x});$$

5: Updating $\boldsymbol{P} = \boldsymbol{P} \cup \{\boldsymbol{p}^*\}$, $\boldsymbol{L} = \boldsymbol{B} \cup \boldsymbol{P}$, and getting the prompt point at location $\boldsymbol{p}^*$;
6: **end for**
7: **return** $\boldsymbol{P}$

2. Expanded Comparative Analyses

We compared the proposed RIM with another line of methods trained using more detailed instance-level annotations on three additional datasets with more categories, *i.e.* ADE20K-847 [14], ADE20K-150 [14], Pascal Context-459 [6]. Among them, ADE20K-150 is a large-scale scene understanding dataset with a total of 150 annotated categories. ADE20K-847 has the same images as ADE20K-150 but encompasses a broader spectrum of annotated classes (847 classes in total), which is the most challenging dataset

*Equal contribution
†Corresponding author

Method	Training Dataset	Supervision label	mask	caption	mIoU A-847	PC-459	A-150	PC-59	PAS-21	COCO
SPNet[9]	Pascal VOC	✓	✓		-	-	-	24.3	18.3	-
ZS3Net[1]	Pascal VOC	✓	✓		-	-	-	19.4	38.3	-
LSeg[4]	Pascal VOC	✓	✓		-	-	-	-	47.4	-
SimBaseline[12]	COCO	✓	✓		-	-	15.3	-	74.5	-
ZegFormer[2]	COCO	✓	✓		-	-	16.4	-	73.3	-
LSeg+[3]	COCO	✓	✓		3.8	7.8	18.0	46.5	-	55.1
MaskCLIP[15]	COCO	✓	✓		8.2	10.0	23.7	45.9	-	-
OSIDE [11]	COCO	✓	✓		11.1	14.5	29.9	57.3	84.6	65.2
SAN [13]	COCO	✓	✓		10.1	12.6	27.5	53.8	94.0	-
GroupViT[10]	GCC+YFCC			✓	4.3	4.9	10.6	25.9	50.7	21.1
OpenSeg[3]	COCO		✓	✓	6.3	9.0	21.1	42.1	-	36.1
OSIDE [11]	COCO		✓	✓	11.0	13.8	28.7	55.3	82.7	52.4
RIM(Ours)	Training-free				6.1	7.8	17.0	34.3	77.8	44.9

Table 1. Open-vocabulary semantic segmentation performance comparison between the proposed training-free RIM and a line of works trained on the COCO dataset, which shares high lable-set similarity with the validation datasets

.

Dataset	Label Sim. to COCO Stuff
ADE20K-847	0.57
Pascal Context-459	0.70
ADE20K-150	0.73
Pascal Context-59	0.86
Pascal VOC	0.91

Table 2. The label-set similarity between validation datasets and training set (*i.e.* COCO Stuff). Measured by Hausdorff distance and cosine similarity based on CLIP text encoder.

for open-vocabulary semantic segmentation. Similarly, Pascal Context-459 has the same image set as Pascal Context-59 but far more annotated classes (459 classes in total).

We conduct an analysis of the relationships between the datasets following SAN [13] by computing the category similarity between other datasets and the training dataset of other methods in Table 2, *i.e.*, COCO Stuff. The similarities are computed with the Hausdorff Distance, and the text embedding of each concept is extracted from pre-trained CLIP [7] text encoder with ViT-L/14 and cosine similarity is employed for pairwise similarity computing. It can be observed from Table 2 that these datasets exhibit a high degree of similarity with the COCO dataset. This implies that methods trained on COCO utilizing detailed instance-level annotations are likely to achieve better performance due to the increased volume of annotation information as shown in Table 1. However, this also correspondingly escalates the training costs to some extent. Our proposed training-free RIM outperforms many methods trained on COCO with pixel-level annotations, demonstrating the open-vocabulary segmentation capabilities of our approach.

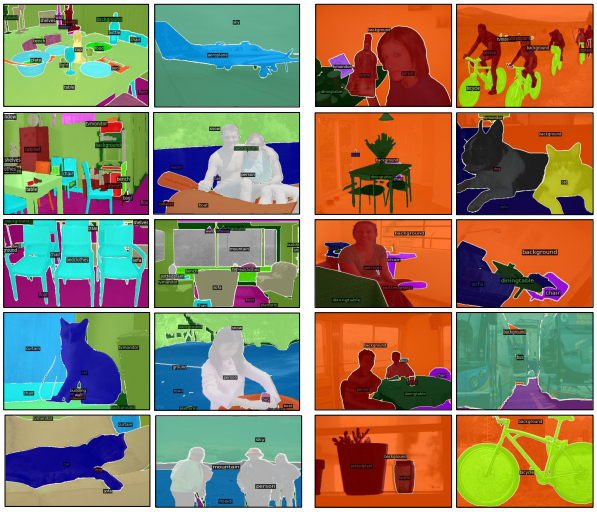

Figure 1. Qualitative results of our method on different datasets.

3. More Visualizations

As illustrated in Figure 1, the visualization of the predictions illustrates that our method enables meticulous concept mining and precise region classification.

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[2] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2

[3] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scal-

ing open-vocabulary image segmentation with image-level labels, 2022. 2

[4] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 2

[5] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8334–8343, 2021. 1

[6] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[9] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 2

[10] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2

[11] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[12] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[13] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 2

[14] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 1

[15] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2